PROJECT MUSE®

Named Entity Normalization: Combining Normalization Rules,
Endogenous Resources and User-Oriented Process /
Normalisation des entités nommées : allier règles
déclaratives, ressources endogènes et processus centré
sur l'utilisateur

Vanessa Andréani, Thibault Roy, Thomas Lebarbé

Canadian Journal of Information and Library Science, Volume 35,
Number 3, September/septembre 2011, pp. 229-263 (Article)

Published by University of Toronto Press

DOI: <https://doi.org/10.1353/ils.2011.0021>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/450027>

Named Entity Normalization: Combining Normalization Rules, Endogenous Resources and User-Oriented Process

Normalisation des entités nommées : allier règles déclara- tives, ressources endogènes et processus centré sur l'utilisateur

Vanessa Andréani, Thibault Roy et Thomas Lebarbé
TecKnowMetrix, 4, rue Léon Béridot, Les Jardins d'Épione,
38500 Voiron, France
LIDILEM – Université Stendhal Grenoble 3, Domaine universitaire,
1180, avenue centrale, 38400 Saint Martin d'Hères, France
va@tkm.fr

Abstract: Normalization is involved in many fields of information processing. It improves the performance of several applications, such as information retrieval or information extraction, and makes the construction of language resources more reliable. Normalization consists in standardizing each variant of a term or named entity into a unique form, and in this way restricts the impact of language variation. Our work applies to named entity normalization, and aims at optimizing fine-grained corpus analyses carried out by the TecKnowMetrix Company. Our approach mixes several methods, such as pattern matching, similarity metrics and endogenous techniques. Moreover, we place the user in the center of our normalization process, in order to obtain fully reliable data that fit his or her needs.

Keywords: normalization, named entities, information processing, information retrieval, corpus analysis, endogenous methods, patterns, similarity metrics, evaluation

Résumé : La normalisation intervient dans de nombreux champs du traitement de l'information. Elle permet d'améliorer l'efficacité d'applications telles que la recherche ou l'extraction d'information, et de rendre plus fiable la constitution de ressources langagières. La normalisation consiste à ramener toutes les variantes d'un même terme ou d'une entité nommée à une forme standard, et permet de limiter les effets de la variation linguistique. Notre travail porte sur la normalisation des entités nommées, et vise à optimiser les analyses de corpus fines réalisées par la société TecKnowMetrix. Notre approche combine plusieurs méthodes, telles que l'utilisation de formes, de calculs de similarité, ou encore de

techniques endogènes. De plus, nous plaçons l'utilisateur au centre du processus de normalisation, afin d'obtenir des données parfaitement fiables et adaptées à ses besoins.

Mots-clés : normalisation, entités nommées, traitement de l'information, recherche d'information, analyse de corpus, méthodes endogènes, formes, mesures de similarité, évaluation

Introduction

Il est possible de réaliser efficacement une recherche d'information ponctuelle sur Internet en soumettant quelques mots clés bien choisis à des moteurs de recherche. Nous en faisons tous quotidiennement l'expérience. Il n'en va pas de même pour des objectifs plus complexes, entre autres dans le domaine de la veille documentaire. Lorsque ce travail de recherche et de veille d'information prend place dans un contexte professionnel, les résultats obtenus doivent être de qualité optimale.

C'est dans un tel contexte que se place l'activité de recherche et développement de la société TecKnowMetrix, qui propose à ses clients des conseils en stratégie de l'innovation. Pour ce faire, elle analyse de grands corpus de brevets, publications scientifiques et articles de presse technico-économique, qui regroupent toute la production scientifique et technique gravitant autour des domaines d'intérêt de nos clients.

L'analyse de ces corpus doit donc être la plus précise possible, et la phase amont de préparation des données est particulièrement importante et délicate. Les documents qui constituent nos corpus d'analyse sont tous associés à des entités nommées, tels les déposants d'une demande de brevet et les inventeurs ou les auteurs de publications et d'articles de presse. Ces entités sont primordiales dans l'analyse de tels corpus, puisqu'elles représentent les acteurs des domaines que nous traitons.

Ces entités nommées sont clairement identifiées dans les corpus que nous analysons. Cependant un important problème subsiste : celui de leur normalisation. En effet, la graphie d'une même entité peut exister de diverses façons. Ces différentes formes graphiques d'une même entité ont une portée considérable au cours de la phase d'analyse.

Dans cet article, nous proposons d'abord de définir les notions liées au problème de la normalisation. Après avoir énoncé différents exemples de

ce problème, nous présentons des solutions existantes et nous identifions leurs limites par rapport à notre problématique. Ensuite, nous détaillons notre processus de normalisation en mettant en avant ses points forts : un double traitement « intra-entité » et « inter-entités » et une place centrale laissée à l'utilisateur. Enfin, nous présentons une évaluation de notre processus en mettant l'accent sur ses forces et ses faiblesses, pour ensuite conclure avec nos perspectives de travail à court terme.

1. Normalisation des entités nommées

1.1. Présentation détaillée du problème

Dans le cadre de l'exploitation de données textuelles pour une analyse fine de l'information, la disparité des formes renvoyant à une même entité nommée ou à un même terme pose rapidement un problème. Par conséquent, il est souvent nécessaire de procéder à la normalisation de ces données avant de commencer l'analyse même. Nous avons donc mis en place un système de normalisation d'entités nommées qui permettra de travailler avec des données non bruitées et fiables.

Nous définissons la normalisation comme un processus qui permet de ramener plusieurs occurrences d'un même type à une forme standard, et dont la finalité est de reconnaître, dans des données structurées ou non structurées, toutes les occurrences d'un même terme ou d'une même entité.

Bien que la normalisation puisse porter sur des noms propres autant que sur des termes « communs », que ce soit notamment des noms, des structures verbales, ou des termes complexes, les travaux présentés portent exclusivement sur la normalisation des entités nommées. Comme Poibeu (2001), nous considérons que les entités nommées sont « *l'ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné* ». L'auteur inclut également dans cette catégorie « *les dates, les unités monétaires, les pourcentages, etc.* ». Parmi tous ces types d'entités nommées, nous ne nous intéresserons qu'aux noms de personnes et d'organisations, puisque ce sont ceux qui apparaissent dans les données que nous souhaitons traiter. Nous distinguons deux sous-catégories d'organisations : les entreprises privées, et les institutions publiques (entre autres les universités, hôpitaux, et instituts).

Le présent travail se place dans une perspective de constitution d'un réseau ontologique ou termino-ontologique d'entités nommées dans un système de recherche et d'extraction d'information. Nous reprenons la définition de Bourigault et al. (2004) qui décrit l'ontologie comme une « *conceptualisation des objets du domaine selon un certain point de vue, imposé par l'application. Elle est conçue comme un ensemble de concepts, organisés à l'aide de relations structurantes, dont la principale, celle avec laquelle est construite l'ossature taxonomique de l'ontologie, est la relation is-a.* ». Comme Mann (2002), nous prenons cette description du concept d'ontologie dans un sens large, et incluons la possibilité de constituer un réseau ontologique sur la base d'entités nommées, et donc de noms propres. Une ressource termino-ontologique est quant à elle un « *modèle de connaissances comportant un réseau conceptuel et des termes associés* » (Aussenac-Gilles 2007). Là aussi, nous considérons que des entités nommées peuvent parfaitement être intégrées à ce type de représentations. Nous partons du principe que les entités nommées, bien qu'elles aient un statut particulier puisqu'elles sont avant tout référentielles parce qu'elles renvoient à un objet réel, n'en sont pas moins représentatives d'une activité ou d'un champ de recherche. À ce titre, elles ont leur place dans une ontologie de domaine ou dans un réseau termino-ontologique.

Or la tâche de normalisation est souvent un préalable utile, voire indispensable, pour plusieurs tâches plus globales, et notamment pour la constitution d'ontologies, la recherche d'information, l'extraction d'information ou l'analyse statistique de corpus.

La normalisation permet d'uniformiser et de rendre fiables les entités ou les termes qui formeront les noeuds du réseau ontologique. Sans cette étape, on risque de fausser le réseau ainsi que toute tâche qui utilise la ressource obtenue. Ainsi, pour la société Mitsubishi¹, la normalisation des occurrences *Mitsubishi Corp*, *Mitsubisi KK* et *Mitsubishi* permet de ramener ces entités à *Mitsubishi* dans le réseau, au lieu d'avoir autant d'entrées censées représenter des organisations différentes que de variantes d'un même nom d'entité.

Pour la recherche d'information, la possibilité d'identifier toutes les formes possibles d'une entité au moyen d'une étape de normalisation permet d'augmenter significativement le taux de rappel. Cette tâche de normalisation présente le même avantage pour l'extraction d'information et les systèmes question-réponse. Par exemple, si la requête ou la question con-

tient *Mitsubisi KK* et le texte *Mitsubishi Corp*, l'appariement pourra avoir lieu entre les deux grâce à la normalisation.

Enfin, pour des analyses statistiques de corpus, si nous voulons identifier et compter tous les brevets déposés par l'entreprise Mitsubishi, une normalisation sera nécessaire afin de ramener diverses occurrences comme *Mitsubishi Corp*, *Mitsubisi KK* et *Mitsubishi* à un seul et même nom d'organisation, *Mitsubishi*.

Du point de vue théorique, la normalisation est souvent intégrée à la tâche d'extraction des entités nommées. Cependant, nous considérons qu'elle doit être traitée à part et d'une manière autonome, bien que dépendante de la tâche dont elle permet d'améliorer les résultats. En effet, la normalisation des entités nommées peut être considérée comme transverse d'un certain nombre d'applications, ainsi qu'on l'a déjà mentionné. Elle interviendra en amont ou en aval du traitement principal en fonction de la tâche globale dans laquelle elle s'insère.

1.2. Solutions existantes

Plusieurs auteurs mentionnent l'existence d'une étape de normalisation dans des systèmes de traitement automatique de langues ou de constitution de ressources linguistiques. Des méthodes de normalisation ont été mises en place pour les graphies de manière générale. Par exemple, Labbé et Labbé (2006) appliquent une méthode de normalisation qui consiste à restituer à chaque mot sa forme normalisée complète, c'est-à-dire non abrégée, débarrassée des majuscules au début de la phrase et des erreurs d'orthographe. Cette forme normalisée est placée dans une étiquette attachée à chaque mot, et accompagnée de son lemme et de sa catégorie grammaticale.

Proche de cette méthode, la racinisation (*stemming*) permet de ramener toutes les variantes morphologiques d'un mot à leur racine commune, ce qui permet notamment d'étendre les requêtes de certains moteurs de recherche, ou encore de définir des motifs d'extraction terminologique en langue du droit (Lebarbé 2007).

Pour les entités nommées, ce mode de normalisation est loin d'être optimal et représente des traitements à la fois mal adaptés et trop peu détaillés. La spécificité des entités nommées empêche en effet une

lemmatisation ou une racinisation à proprement parler, et demande d'autres traitements plus adéquats. D'autre part, une lemmatisation ou une racinisation empêche le traitement correct des noms composés : ce type de technique portant sur les mots graphiques, un enchaînement de deux ou plusieurs mots graphiques formant un mot composé, ou un terme ne sera pas traité comme tel.

Il existe donc de nombreuses méthodes de normalisation adaptées aux entités nommées dans le cadre des systèmes de traitement de l'information. Ces techniques de normalisation peuvent différer selon le type de données en entrée, d'approche adoptée, et d'entités nommées traitées, et l'objectif final de la tâche globale. Dans tous les cas, le but est de ramener une occurrence d'entité nommée à une forme standard, ou pour le moins de rassembler toutes les occurrences différentes d'une même entité.

Nous présentons des systèmes de normalisation représentatifs des approches possibles :

- [...] approches qui s'appuient sur des ressources externes;
- [...] méthodes fondées sur des formes d'extraction;
- [...] cas particulier des données structurées.

Approches prenant appui sur des ressources externes

Pour la normalisation d'entités nommées dans leur acception la plus restrictive, comme les noms propres de lieux, d'organisations ou de personnes, une approche courante consiste à utiliser un dictionnaire ou une ressource externe suffisamment exhaustive, de manière à ramener chaque occurrence d'une entité à l'entrée « standard » qui lui correspond dans la ressource. On peut ainsi résoudre les difficultés liées à la variation linguistique, notamment en termes de synonymie. Le problème est traité par Khalid et al. (2008) dans le cadre d'une recherche d'information pour un système question-réponse, sur un corpus d'articles de presse et de fils RSS. Par la normalisation, les auteurs cherchent à régler deux difficultés majeures pour ce type de tâche : le problème de l'ambiguïté d'une entité comme *Alabama* qui peut désigner la rivière ou l'état américain, et celui de la synonymie, évoqué plus haut, lorsque plusieurs noms réfèrent à la même entité, par exemple *America* et *USA*. Les auteurs résolvent ces deux aspects problématiques en utilisant Wikipédia comme ressource externe :

pour la gestion de la synonymie, ils utilisent les liens de redirection du site, et pour les noms qui peuvent renvoyer à plusieurs entités, ils emploient les pages de désambiguïsation. En comparant les résultats obtenus par le système avec et sans normalisation en amont, ils démontrent que la tâche de recherche d'information est plus efficace lorsque les entités nommées ont été préalablement normalisées.

Dans Jijkoun et al. (2008), la même équipe de recherche propose une méthode similaire de normalisation des entités nommées, fondée sur l'exploitation de Wikipédia, mais cette fois adaptée aux contenus générés par l'utilisateur, comme des blogs ou des commentaires de lecteurs sur des articles. Là encore, ils constatent que les sources d'erreur les plus importantes du travail de normalisation sont la synonymie et l'ambiguïté de certaines entités nommées. L'un des points forts de leur méthode est qu'elle est adaptable à plusieurs langues, pourvu que les entités normalisées soient répertoriées dans Wikipédia.

Ces méthodes sont efficaces sur du texte tout-venant, comme les corpus de presse ou de blogs. Nous pouvons raisonnablement supposer que les entités nommées présentes dans ces textes relèvent de connaissances du monde d'ordre général, et c'est pourquoi Wikipédia est très bien adapté dans ce contexte. Cependant, dès que les domaines abordés sont très spécialisés ou que les entités nommées sont, par exemple, des noms de petites entreprises, il est très probable que ce type de ressource et les méthodes dans lesquelles il s'insère soient insuffisants.

Méthodes fondées sur des formes d'extraction

Dans les systèmes d'extraction d'information de domaines spécialisés, l'utilisation de formes pour la normalisation permet de régler efficacement les problèmes de variation linguistique. Encore une fois, ces formes sont applicables aux documents non structurés. Selon les systèmes, l'acception du terme *entité nommée* peut être assez large et inclure des éléments autres que les entités nommées courantes. Alphonse et al. (2004) travaillent, dans le cadre du projet Caderige, à l'extraction d'information à partir de données prélevées dans des textes du domaine biomédical, et plus particulièrement des interactions géniques. Dans un premier temps, ils normalisent des noms de gènes et de protéines, en l'occurrence considérées comme des entités nommées. Des amorces de synonymie comme « *formerly* » ou « *also called* » et des formes comme « *gene* amorce *gene* » permettent de gérer la variation linguistique. Ces

formes sont ensuite entraînées sur des échantillons de données, puis projetées sur les données de manière à les normaliser. La normalisation atteint une précision de plus de 97 %, et un rappel de 75 %. En outre, les auteurs démontrent que cette étape de normalisation « *facilite l'acquisition et l'apprentissage de règles d'extraction en offrant une représentation plus abstraite des phrases* ». Avec le même objectif, mais dans le domaine financier cette fois, Poibeau (2003) prône lui aussi une phase de normalisation des entités nommées préalable à l'extraction, et la qualifie d'*étape [...] essentielle*. C'est principalement la nature des entités nommées qui change ici, puisqu'il s'agit pour la plupart de noms d'organisations et de personnes, de montants et de dates, et non plus de gènes.

Une méthode de normalisation par formes est pertinente et remplace efficacement les dictionnaires et autres ressources pour le traitement de documents spécialisés. Bien que nos données soient différentes en termes de domaine comme en ce qui a trait à leur structure, le principe de la normalisation par formes est approprié à notre travail, du moins en partie.

Cas particulier de données structurées

Les documents non structurés ne sont pas les seuls à nécessiter une normalisation. Normaliser des données structurées peut être nécessaire lorsqu'elles sont bruitées ou incohérentes dans leur version brute, et si elles réduisent l'efficacité des traitements lors de leur entrée. L'existence de doublons dans une même base de données ou dans plusieurs bases différentes pose un problème lorsque la forme des entités nommées varie, ce qui rend tout recoupement impossible. Les méthodes relèvent alors souvent de calculs de similarité entre champs, de manière à rapprocher deux ou plusieurs noms qui font référence au même objet du monde. Elmagarmid et al. (2007) recensent les méthodes utilisées pour la détection d'entrées dupliquées dans les bases de données structurées. Cette détection consiste en grande partie à normaliser les entrées en question. Cela permet d'apporter une réponse aux difficultés posées par l'hétérogénéité lexicale des bases de données.

Les calculs de similarité ont l'avantage d'être peu dispendieux en ce qui a trait à la mise au point, et offrent des résultats satisfaisants pour certains aspects de la normalisation, en particulier pour la correction des erreurs typographiques ou orthographiques. L'inconvénient de ces techniques réside dans leur coût computationnel.

Elmagarmid et al. (2007) relèvent un nombre relativement important de techniques de mise en correspondance entre champs. La plupart d'entre elles s'appuient sur des mesures de similarité entre caractères, mots graphiques ou phonèmes, grâce auxquelles il est possible d'établir des équivalences entre les noms et de produire ainsi une liste de variantes pour une même entité nommée.

1.3. Limites de ces approches

L'ensemble des techniques et approches mentionnées permet d'obtenir des données relativement fiables et non bruitées, et d'améliorer l'efficacité des tâches qui s'appuient sur ces normalisations. Cependant, nous avons pu constater que la plupart de ces systèmes de normalisation portent sur des données bien déterminées, avec des types d'entités nommées parfois très précisément circonscrits, et avec des objectifs très clairement définis. La construction d'un système de normalisation dépend du but visé. Dans l'état actuel des avancées théoriques et techniques, dès que l'un des paramètres change, le système peut devenir mal adapté. Par exemple, les méthodes utilisées pour la normalisation d'entités nommées dans des documents non structurés sont parfois inutiles pour la normalisation de champs de bases de données, et inversement. Citons le cas du repérage de formes lexico-syntaxiques précises comme « *gene* amorce *gene* », où « amorce » devient un adverbe ou une conjonction, très utile dans du texte comme chez Alphonse et al. (2004), mais inefficace dans des champs de base de données. Dans le meilleur des cas, les champs ne contiennent que des syntagmes ou groupes nominaux, avec peu d'adverbes, de prépositions ou de conjonctions. À l'inverse, les mesures de similarité utilisées pour la normalisation des bases de données (Elmagarmid et al. 2007) sont très mal adaptées à la gestion de la synonymie fortement présente dans les textes (Alphonse et al. 2004).

Si l'adaptabilité d'un système entre différents types de données est difficilement envisageable, il en va de même pour ce qui relève de la transportabilité entre domaines : il est évident qu'en fonction du domaine envisagé, les formes de repérage utilisées doivent être modifiées, voire supprimées pour en créer d'autres plus pertinentes. En effet, une forme du type Alphonse et al. (2004) devient inutile si l'entité recherchée est une organisation dans un corpus de textes traitant de finance et non plus un gène dans des textes biomédicaux. De même, l'utilisation d'une ressource externe comme Wikipédia, comme chez Jijkoun et al. (2008)

et Khalid et al. (2008), devient vite insuffisante, voire génératrice de parasites, dès que le système doit traiter un domaine très spécialisé, comme la physique nucléaire ou les biocarburants.

Si nous nous intéressons au cas particulier des techniques de détection de doublons en base de données, nous constatons qu'il existe deux grands types de méthodes, les mesures de similarité et les techniques d'apprentissage. Chacun de ces types présente des avantages pour la normalisation d'entités nommées, mais comporte aussi ses propres limites. Les mesures de similarité, qui sont des calculs extrêmement efficaces pour certaines données, peuvent être insuffisantes dès que les entrées à rapprocher sont du type :

1. Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA.
jowell@gx.net
2. University of Kentucky, Lexington 40536-0098, USA.
runge@pop.uky.edu

Dans ce cas, l'objectif est de ramener ces deux entrées à *University of Kentucky*, et il semble difficile d'y parvenir en utilisant uniquement des mesures de similarité.

Les techniques d'apprentissage, quant à elles, ont souvent besoin d'un corpus d'entraînement, qui doit être constitué et étiqueté manuellement. Mais cette constitution manuelle est la plupart du temps très dispendieuse.

Il est à noter que l'utilisateur n'a pas un rôle central dans les méthodes que nous venons de présenter. Or, il nous semble primordial que l'utilisateur expert intervienne dans le processus de normalisation, de manière à rendre les données fiables et adaptées à l'usage auquel elles sont destinées. Bien que les techniques d'apprentissage requièrent la constitution manuelle de données d'entraînement, nous n'assimilons pas cette phase à une validation par l'utilisateur expert, pour deux raisons :

- ce n'est pas une validation des données traitées, mais un travail en amont permettant l'exécution du processus;
- elle n'est pas systématiquement réalisée par un expert du domaine.

Chacune des méthodes s'articule autour d'une seule approche, comme l'utilisation d'une ressource externe pour Jijkoun et al. (2008) et Khalid et al. (2008), ou des formes de repérage pour Alphonse et al. (2004). Une méthode de normalisation combinant plusieurs approches, avec des traitements à différents niveaux textuels, nous semble donc particulièrement pertinente. Nous sommes d'avis qu'il est judicieux de tirer parti à la fois des informations contenues dans l'entité nommée même à normaliser (approche « intra-entité »), et de celles que peut fournir une comparaison de cette entité avec les autres entités du corpus (approche « inter-entités »). De manière générale, on peut ainsi construire des systèmes plus adaptables en fonction des données, et plus transportables d'un domaine à un autre. Dans la perspective de notre travail, une approche mixte nous donne la possibilité de résoudre les divers types de problèmes.

2. Processus interactif et itératif pour la normalisation

Nous avons présenté le problème que posaient des données textuelles non uniformes et les solutions existantes en matière de normalisation en fonction de l'application finale. Ces solutions peuvent être limitées, notamment en termes d'adaptabilité. Nous proposons donc une solution, d'abord par un exposé du contexte dans lequel elle intervient, puis par l'explicitation du processus à partir de plusieurs exemples.

Notre traitement est effectué sur des données structurées, et plus précisément sur une base de données dans laquelle une table contient l'ensemble des organisations qui déposent un brevet ou publient un article scientifique. De fait, les entités nommées sont déjà clairement identifiées. Le problème vient de la disparité des formes pour des entités qui réfèrent à un même objet du monde, mais aussi de la coprésence de plusieurs sous-entités dans un seul et même nom d'organisation. Notre objectif est de normaliser les entités nommées pour qu'elles correspondent aux attendus définis en amont avec les utilisateurs experts. Nous tenons donc compte de leurs attentes, et non de normes de représentation existantes comme les normes AFNOR. De plus, afin de répondre aux attentes des utilisateurs, la normalisation en sortie du traitement est monolingue, alors que nos données sont multilingues.

Dans cette partie, nous caractérisons notre corpus de travail, puis nous présentons le processus de normalisation qui lui est appliqué.

2.1. *Caractérisation du corpus*

La base de données interne de l'entreprise TecKnowMetrix comprend plusieurs millions de brevets et articles scientifiques, que nous regrouperons sous le terme de publications. Toutes ces publications relèvent de domaines scientifiques et techniques, dont les principaux sont, entre autres, la médecine, la chimie, la biologie, et l'électronique. Toutes les informations relatives à chaque publication sont stockées dans des tables liées à celles contenant le corps de la publication. Ces informations sont notamment le titre, la source de la publication, le déposant s'il s'agit d'un brevet ou encore l'auteur s'il s'agit d'un article.

Une table contient les données relatives au déposant ou à l'auteur des publications de la base. Elle contient entre autres le nom et l'adresse de l'organisation. Elle est actuellement composée de plusieurs millions d'entrées dont le nombre augmente régulièrement. C'est cette table qui forme notre corpus de travail, et plus précisément le champ « nom d'organisation » de chaque entrée. Il correspond au nom de l'organisation tel qu'il a été récupéré lors de l'import de la publication. Les sources à partir desquelles les informations sont importées sont assez nombreuses, ce qui entraîne une grande diversité des conventions graphiques.

2.2. *Présentation globale du processus*

Étant donné les difficultés évoquées à la section 1.3, nous avons adopté une approche mixte et combiné des traitements à base de règles, des méthodes de mesures de similarité et des méthodes d'extraction endogène. Nous nous plaçons aussi bien du point de vue de l'intra-entité lors de l'application des règles, que de l'inter-entités, lorsqu'il s'agit des calculs de distance d'édition et des méthodes endogènes. De plus, nous intégrons l'utilisateur dans le processus, car nous considérons qu'il est le seul qui peut déterminer la pertinence et la cohérence de certaines normalisations. Il a donc la possibilité d'interagir avec le système en validant ou invalidant certaines propositions. En fonction des choix de l'utilisateur, le système s'adapte de manière à ce que les validations déjà acceptées ne lui soient plus proposées.

Le processus prend en entrée le nom d'organisation tel qu'il a été importé de la source d'information. En sortie, il restitue ce nom de départ et lui associe un nom normalisé. Les informations d'origine sont ainsi conservées malgré les transformations.

Le choix de ces traitements se justifie par les types de problèmes rencontrés dans notre corpus. Nous pouvons classer les difficultés majeures en trois catégories distinctes et relativement indépendantes :

- les erreurs typographiques et orthographiques courantes, comme pour *Mitsubisi* au lieu de *Mitsubishi*;
- la diversité des sources dont les données du corpus proviennent, ce qui entraîne une grande disparité dans la graphie et dans la description d'une même entité, par exemple, *Mitsubishi Corp* et *Mitsubishi*;
- pour les entités nommées désignant des institutions publiques en particulier, il n'est pas rare de trouver à l'intérieur d'un même nom d'organisation plusieurs sous-éléments d'une même entité : [...]

Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA. jowell@gx.net

est assez représentatif. Il devient alors difficile de dégager l'entité de niveau supérieur en mettant de côté les sous-organisations qui viendraient bruyter les données.

La figure 1 montre le déroulement du processus de normalisation dans son ensemble. Par la suite, nous décrivons en détail chaque étape de la méthode.

La phase de délimitation des entités nommées permet, grâce à des ensembles de règles et à des lexiques détaillés à la section 2.3.1, de découper un nom d'organisation brut pour en extraire la partie la plus « générale » et la plus appropriée à l'application finale. De plus, elle procède à la réécriture, dans une forme standard, de chaque entité nommée ainsi extraite.

Grâce à des mesures de similarité, la phase 2 rapproche des noms d'organisations ayant en commun un certain nombre de caractères ou de graphies. Le rapprochement s'effectue automatiquement, puis l'expert corrige les erreurs et valide les appariements.

La phase 3 est fondée sur une technique endogène et elle permet de régler des cas non résolus à la fin de la phase 2. En tant que technique endogène, elle s'appuie sur les données du corpus à traiter lui-même. Là

Figure 1 : Processus global de normalisation

encore, chaque normalisation effectuée à ce stade est soumise à la validation de l'utilisateur, ce qui garantit la fiabilité des données.

Enfin, les nouvelles normalisations sont stockées dans la base de données; chaque nouveau processus de normalisation s'appuiera donc sur les données entrées au fur et à mesure.

2.3. Processus détaillés

2.3.1. Phase 1 : Délimitation et réécriture des entités nommées—traitement intra-entité

La phase 1 (n° 1 dans le schéma), intervient immédiatement après l'importation des données à normaliser dans les bases de données internes. Elle consiste à appliquer une ou plusieurs règles de pré-normalisation tirées d'un ensemble de règles. Les règles appropriées s'appliquent automatiquement lorsque le système détecte des déclencheurs. Ces déclencheurs sont des termes ou des structures syntaxiques qui indiquent la présence d'une

entité nommée dont la forme doit être modifiée pour correspondre aux normes que nous avons définies. De plus, certaines règles permettent de typer l'organisation traitée. Elle peut être identifiée comme une entreprise, une institution publique, ou encore comme un particulier quand un individu a déposé une demande de brevet ou publié un article en son nom propre. Enfin, des règles sont appliquées pour déterminer le pays d'origine de l'organisation examinée.

Nous avons constitué trois lexiques sur lesquels s'appuient les ensembles de règles pour normaliser les entités et extraire les renseignements complémentaires que sont le type et le pays d'origine. Ces listes sont placées au début du système de règles :

- un lexique d'amorces;
- une liste de mots vides dans les cinq langues les plus présentes dans le corpus;
- une liste des pays du monde.

Le premier lexique est une liste hiérarchisée des amorces permettant de détecter la présence, le type et parfois la nationalité d'une organisation. Ces amorces peuvent être des identifiants nationaux de sociétés, tels que *GmbH*, *Corporation* ou *S.A.*, ou des noms génériques d'entités, comme *University*, *Institute* ou encore *Center*. Pour chaque amorce, les formes développées comme les formes abrégées sont prises en compte. Nous considérons que cette liste est relativement fermée, puisque ces amorces sont en quelque sorte des marqueurs morphologiques qui « signalent » l'instanciation d'un type donné d'entité. À ce titre, si le nombre d'entités nommées est potentiellement infini, celui des marqueurs de type d'entité est quant à lui relativement restreint. Notre liste, dont nous présentons ici un extrait, est constituée d'une centaine d'articles.

```
ETS S\.? ?N\.? ?C\.?France
AC1 UNIVunknown
AC2 Ospedale Italy
AC5 Dep(arte?ment|t)?\.? unknown
```

Cette liste a été constituée de manière à pouvoir être interprétée par notre programme. La première partie de chaque ligne, *ETS* ou *AC1* / *AC2* / ..., permet d'identifier le type d'amorce. *ETS* correspond aux

entreprises, tandis que AC renvoie aux institutions publiques. Le chiffre qui suit AC précise l'endroit où se situe l'amorce dans la hiérarchie : les universités sont les institutions publiques du plus haut niveau. Le deuxième article correspond à l'amorce même. Pour la première amorce, les occurrences *SNC*, *S N C*, *S.N.C.*, *S. N. C.*, etc. sont toutes repérées et identifiées. La dernière amorce permet de repérer toutes les variantes de *Department*, comme *Dep.*, *Dept.*, etc. Enfin, le dernier article précise le pays auquel l'amorce détectée peut être associée de manière certaine. Ainsi, si le système rencontre l'amorce *ospedale*, une règle lui permet de déduire que le pays d'origine de cette organisation est l'Italie.

Le deuxième lexique contient une liste de mots vides au sens de Tesnière (1969), qui sert à déterminer certaines constructions syntaxiques utiles à la normalisation. Ils permettent en effet, en combinaison avec le lexique d'amorces, de construire les formes d'extraction et de normalisation utilisées dans les règles. Par contre, les mots vides sont éliminés à partir de cette liste à la fin de la première étape de normalisation. Cette liste regroupe les mots vides utilisés dans les données en cinq langues, de manière à couvrir le plus d'occurrences possible dans le corpus. Les langues les plus présentes sont l'anglais, l'allemand, l'espagnol, le français et l'italien. Notre lexique est constitué d'une quarantaine d'entrées, qui présentent la catégorie grammaticale du mot-outil, le mot-outil lui-même, et enfin sa langue.

Enfin, une liste de pays a été constituée pour identifier le pays d'origine d'un nom d'organisation, lorsque ce dernier est explicitement cité. Elle est uniquement composée de noms de pays en anglais, puisque lorsque le pays d'origine est cité dans les noms d'organisations, il l'est presque exclusivement dans cette langue.

Ces trois lexiques sont les seules ressources externes que nous utilisons. En effet, nous avons décidé de solliciter le moins possible de ressources de type dictionnaire ou thésaurus, dans la mesure où nous considérons comme impossible, ou à tout le moins trop dispendieux, d'avoir un dictionnaire de noms propres suffisamment important pour couvrir toutes nos données, surtout dans des domaines spécialisés (Poibeau 2001).

À l'exception de la liste des pays, ils ont été construits de manière empirique, à partir d'observations faites en corpus.

Les règles se répartissent en deux groupes :

- les règles de sélection du segment le plus pertinent;
- les règles de normalisation du segment retenu.

Sélection du segment le plus pertinent

Le premier jeu de règles permet de sélectionner la partie du nom d'organisation brut qui contient le nom d'organisation final.

a) Découpage des noms

Le cas échéant, la première de ces règles découpe chaque nom d'organisation aux virgules. Dans la plupart des cas, chaque segment entre virgules renvoie à une sous-entité d'une même organisation. Dans les noms d'organisations :

3. Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia19104.
4. Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA.
jowell@gx.net

nous constatons que chaque segment entre virgules renvoie à un objet du monde intégré à celui que dénote la séquence suivante : en 3. par exemple, le département de radiologie fait partie de l'hôpital universitaire de Pennsylvanie, qui se trouve dans la ville de Philadelphie.

De cette manière, nous pouvons nous assurer que chaque segment est relativement cohérent et autonome, bien qu'il soit hiérarchiquement lié aux autres. Nous verrons par la suite que cette segmentation n'est pas toujours parfaitement adéquate. Toutefois, elle permet un premier découpage grossier, qui s'avère souvent suffisant.

Chaque segment ainsi isolé est considéré comme contenant potentiellement le nom d'organisation final. Si aucune virgule n'est présente, c'est la totalité du nom d'organisation qui est sélectionnée et envoyée directement au début du deuxième jeu de règles. Ce sera le cas pour le nom d'organisation :

5. *Mitsubisi KK*

b) Repérage des amorces

Dans les cas où plusieurs segments ont été identifiés, la deuxième règle sélectionne ceux qui contiennent une amorce faisant partie de la liste. Si aucune amorce n'est trouvée, la troisième règle tente de détecter une suite de majuscules. En effet, ces suites de majuscules au sein de noms saisis pour la plus grande partie en minuscules sont dans la plupart des cas des acronymes d'organisations. En cas de réussite, c'est le segment contenant cette suite de majuscules qui est conservé. En cas de nouvel échec, le système envoie l'ensemble du nom d'organisation brut au début du groupe de règles suivant. Pour nos exemples 3. et 4., les segments retenus comme candidats sont donc les suivants :

3. *Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia19104.*

3 a) *Department of Radiology*

3 b) *Hospital of the University of Pennsylvania*

4. *Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA. jowell@gx.net*

4 a) *Division of Gynecologic Oncology*

4 b) *Department of Obstetrics and Gynecology*

4 c) *University of Kentucky Medical Center*

c) Sélection du segment pertinent

La dernière règle du premier groupe intervient dans les cas où plusieurs candidats ont été retenus, c'est-à-dire lorsque le système a détecté plusieurs amorces réparties dans plusieurs segments. Elle « classe » ces candidats en fonction du niveau hiérarchique de l'amorce qu'ils contiennent. Par exemple, un candidat contenant *University* sera sélectionné face à un autre segment contenant *Department*. Ainsi, pour 3. et 4., les segments considérés comme les plus hauts dans la hiérarchie sont :

3. *Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia19104.*

3 b) *Hospital of the University of Pennsylvania*

4. *Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA. jowell@gx.net*

4 c) *University of Kentucky Medical Center*

puisque, dans chaque cas, *University* est l'amorce du niveau hiérarchique le plus haut.

À la suite des traitements effectués par le premier jeu de règles, le segment contenant le nom d'organisation du niveau hiérarchique le plus haut a donc été identifié. Si le résultat est indécis, les segments de niveau hiérarchique équivalent sont tous conservés.

Normalisation du segment retenu

Le deuxième jeu de règles a pour but de normaliser en soi le candidat retenu, c'est-à-dire de le nettoyer et d'en réécrire certaines parties au besoin afin qu'il corresponde aux normes que nous avons définies.

a) Second découpage

Il arrive que le premier découpage aux virgules ne soit pas suffisamment fin, et un segment placé au début du deuxième ensemble de règles peut contenir plusieurs amorces de niveaux hiérarchiques différents (*cf. supra*). En 3 b) et 4 c), nous trouvons en effet deux amorces différentes dans le même segment :

3. *Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia*19104.
3 b) **Hospital** of the **University** of Pennsylvania
4. *Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA. jowell@gx.net*
4 c) **University** of Kentucky Medical **Center**

Dans un tel cas, une règle est appliquée pour tenter un second découpage, à l'intérieur même du segment. Elle s'appuie sur des formes permettant de repérer des structures qui autorisent une segmentation claire en syntagmes nominaux, et donc une identification sans équivoque du nom d'organisation le plus cohérent. Nous nous appuyons notamment sur la structure :

prep (+ article)? + amorce_de_niveau_supérieur

En isolant le syntagme contenant l'amorce de plus haut niveau, cette règle permet de déterminer de façon certaine que tout ce qui intervient

avant l'amorce n'appartient pas au nom d'organisation hiérarchiquement supérieur, et peut donc être éliminé. Ainsi en 3 b), nous sommes en mesure d'affirmer que le sous-segment *Hospital of the* peut être écarté, pour ne conserver que :

3. *Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia*19104.
 3 b') *University of Pennsylvania*

Cependant, de telles structures ne sont pas systématiques, et un découpage univoque entre des groupes ou des syntagmes n'est pas toujours possible. Ainsi, certains cas restent problématiques même après cette phase du traitement, comme en 4 c), pour lequel il est impossible de déterminer à l'aide de formes et sans ressource externe exhaustive où effectuer la segmentation, de manière à isoler *University of Kentucky*.

b) *Réécriture des noms d'organisations*

Les traitements effectués ensuite sur le segment diffèrent en fonction du type d'amorce identifiée. S'il s'agit d'une amorce de type *entreprise*, elle sera supprimée, de manière à pouvoir identifier comme une seule et même entreprise toutes les filiales d'une maison mère, sous réserve que les filiales aient toutes le même nom. Le nom d'organisation 5. subira donc la transformation suivante :

5. *Mitsubisi KK*
 5 a) *Mitsubisi*

En revanche, si l'amorce relève du type *Institution publique*, elle ne sera pas supprimée, puisqu'elle permet de distinguer plusieurs sous-catégories d'institutions, comme *Hospital*, *Institute* ou *University*. Elle sera réécrite systématiquement, et parfois déplacée en tête de segment, notamment pour les universités, de manière à ce que tous les noms d'organisation relevant de la même sous-catégorie *Université* soient identifiables immédiatement. Notons que ce traitement est appliqué même dans le cas de 4 c), en dépit du fait que le découpage n'ait pu être effectué correctement à ce stade. Ainsi, 3 b') et 4 c) deviennent respectivement :

3. *Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia*19104.
 3 b'') *Univ of Pennsylvania*

4. *Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA. jowell@gx.net*
4 c'') *Univ of Kentucky Medical Center*

c) *Nettoyage final*

En dernier lieu, le segment ainsi transformé est nettoyé : les mots vides et la ponctuation sont supprimés, et certains parasites, comme les adresses de courrier électronique par exemple, sont éliminés.

Pour nos trois exemples, les segments récupérés en sortie de la phase 1 (état *a* des données sur la figure 1) sont donc :

3. *Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia19104.*
3 b'') *Univ Pennsylvania*
4. *Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Kentucky Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA. jowell@gx.net*
4 c'') *Univ Kentucky Medical Center*
5. *Mitsubisi KK*
5 a) *Mitsubisi*

2.3.2. Phase 2 : Appariement par mesures de similarité—traitement inter-entités

À la fin de la phase 1, soit dans l'état *a* des données (Figure 1, p. 242), nous obtenons des noms d'organisations pré-normalisés, dont la majorité sont satisfaisants. Cependant, certains cas ne sont pas résolus. En effet, nos traitements « intra-entité » atteignent ici leurs limites. Entre autres, notre système de règles ne tient pas compte des occurrences de noms d'organisations qui contiennent des erreurs de typographie ou des fautes d'orthographe, c'est pourquoi notre exemple 5 a) n'a pas encore sa forme correcte, qui serait :

5 a') *Mitsubishi*

au lieu de :

5 a) *Mitsubisi*

Pour résoudre ce problème, nous avons convenu d'utiliser des mesures de similarité, en deux étapes. Tout d'abord, nous avons mis en place un calcul de distance d'édition qui permet de comparer entre elles deux chaînes de caractères. Ce calcul est représenté par la phase 2 du processus (Figure 1). La distance de Levenshtein (1966) est donc appliquée pour chaque nom pré-normalisé, qui est comparé à tous les autres noms de la table : le nombre d'insertions, de suppressions et de substitutions de caractères nécessaires au passage d'une chaîne à une autre est calculé. Cette distance est pondérée par sa mise en rapport avec la longueur en caractères de la chaîne la plus importante du couple. Lorsque la distance obtenue est faible, nous considérons que nous avons de fortes chances d'être face à un seul et même nom, dont l'une des occurrences comporte une erreur de typographie ou une faute d'orthographe. Chaque couple de noms ayant une distance faible est ensuite proposé à l'utilisateur (étape *b* du schéma), qui peut valider le rapprochement des deux entités, ou au contraire signifier au système qu'il s'agit de deux organisations distinctes malgré leur degré de similarité élevé (Figure 1, processus 2').

Les choix de l'utilisateur sont conservés en mémoire, de manière à pouvoir être réutilisés si les mêmes cas se présentent lors d'une future normalisation. Ainsi, si la base de données dans laquelle se trouve le nom d'organisation 5 a) contient aussi une entrée dont le nom d'organisation est *Mitsubishi*, alors le couple (*Mitsubisi*, *Mitsubishi*) sera proposé à l'utilisateur, qui n'aura plus qu'à sélectionner *Mitsubishi* comme nom correct : cela aura aussi un effet sur tous les noms d'organisations ayant la même erreur de typographie qu'en 5 a). Le résultat de ces traitements est représenté en *b* sur notre schéma (Figure 1).

Cette validation de l'expert permet de nous assurer que les entités rapprochées le sont à bon escient, une sécurité dont nous n'aurions pas bénéficié si nous avions procédé de manière tout à fait automatique. Nos observations en corpus lors des tests nous ont en effet permis de constater que certains couples d'entités nommées, en dépit de distances très faibles, référaient à des organisations distinctes.

À terme, nous avons l'intention de mettre en place une deuxième mesure de similarité, qui porterait cette fois sur l'unité mot graphique, ou du moins sur une chaîne de caractères, et non plus sur l'unité caractère. Notre objectif est de résoudre des difficultés posées par des noms d'organisations qui, s'ils ne contiennent pas forcément d'erreur de typographie, n'en sont pas moins problématiques, soit : [...]

6. *Mitsubishi*
7. *Mitsubishi Europe*
8. *Mitsubishi Chem*
9. *Mitsubishi Denki*

Dans tous ces cas, nous avons différentes agences ou filiales de la même entreprise, et souhaitons donc les normaliser toutes de la même manière, soit :

10. *Mitsubishi*

Or une distance d'édition calculée sur les caractères serait mal adaptée au rapprochement de toutes ces occurrences. Plusieurs possibilités s'offrent à nous en matière de mesures de similarité. Une première solution serait d'adapter la distance de Levenshtein à l'unité mot graphique : au lieu de compter le nombre de caractères qu'il faut modifier pour passer d'une chaîne à l'autre, il est possible de compter le nombre de mots graphiques qu'il faut effacer, insérer ou substituer pour aboutir à la chaîne *c2* à partir de la chaîne *c1*. De cette manière, nous pourrions rapprocher les noms d'organisations 6. à 9., et les proposer à l'utilisateur pour qu'il avalise la normalisation en *Mitsubishi*. Le problème majeur posé par cette méthode serait celui des mots très fréquents dans l'ensemble du corpus, tels que le mot *Europe* en 7. : rien ne garantit que deux entités nommées contenant le mot *Europe* représentent la même organisation. De fait, nous pouvons nous attendre à un taux de bruit relativement élevé en nous contentant de cette méthode. D'autres techniques intéressantes sont mentionnées par Elmagarmid et al. (2007) et offrent des solutions de remplacement.

Comme pour les données obtenues par les calculs de distance d'édition sur les caractères pour corriger les erreurs typographiques, les résultats obtenus lors de cette phase sont présentés à l'utilisateur pour validation. Les normalisations suivantes s'appuient sur ces données validées.

2.3.3. Phase 3 : Appariement endogène—traitement inter-entités

Cette étape du processus (Figure 1, phases 3 et 3') s'applique seulement à certaines occurrences du corpus : seuls les noms pré-normalisés qui contiennent encore plusieurs amorces de niveaux hiérarchiques différents à la fin de la phase 1 y sont traités. Rappelons les trois noms d'organisations que nous obtenons à la fin de la phase 2 :

3 b'') *Univ Pennsylvania*

4 c'') *Univ Kentucky Medical Center*

5 a') *Mitsubishi*

Le nom 5 a' ne pose aucune difficulté de cet ordre, comme nous l'avons vu en phase 1. En 3 b'', nous avons vu que la phase 1 permettait de régler le problème, et d'extraire correctement *Univ Pennsylvania* grâce au repérage de structures syntaxiques. En revanche, en 4 c'', le problème reste entier, puisque aucune structure syntaxique ne nous permet de découper ce nom en syntagmes de manière univoque, de façon à isoler le nom d'organisation le plus cohérent. Nous ne pouvons nous contenter de nous appuyer sur les mots outils, car l'anglais en utilise moins que le français. Or, la plupart des cas problématiques à ce stade sont des noms en anglais : cela s'explique par le fait que, par exemple, le français utilise fréquemment des modificateurs de nom de la forme *de_Syntagme Nominal* venant se placer après le nom modifié, l'anglais utilise souvent comme modificateurs des noms adjectivés. Ainsi, là où nos formes auraient permis de découper une suite brute comme *Centre médical de l'université du Kentucky* grâce à la présence du mot vide *de*, rien n'indique la « frontière » pour *University of Kentucky Medical Center*, où *University of Kentucky* est le nom adjectivé qui joue le rôle de modificateur.

L'extraction endogène peut être un moyen de déterminer la structure d'un nom d'organisation. Un certain nombre de recherches, telles que Bourigault (1993), Bourigault et Frérot (2006) et Vergne (2004), font état de techniques endogènes, qui représentent un moyen de résoudre des cas où des ressources externes ou des méthodes à base de règles et de formes atteignent leurs limites.

Une approche endogène fondée sur la récurrence de séquences nous paraît tout à fait appropriée en ce qui a trait à nos données et à notre objectif. Nous suivons en cela Vergne (2004), qui expose une méthode d'analyse syntaxique partielle sans ressources de corpus multilingue. Cette méthode s'appuie entre autres sur la longueur et la fréquence des mots. Bien que notre objectif diffère de celui de l'auteur, nous pouvons reprendre un certain nombre de principes et les adapter à notre travail.

En effet, en nous fondant sur les fréquences de segments de différentes longueurs, et non plus sur des structures syntaxiques ou des indices lexicaux, nous contournerons plusieurs écueils. Tout d'abord, cela évite

d'avoir à constituer des lexiques adaptés, ce qui serait très dispendieux puisque nous travaillons sur des données multilingues. Pour les mêmes raisons, mettre en place des formes syntaxiques couvrant tous les cas pour chaque langue semble quasi impossible. D'autre part, du fait que nous traitons exclusivement des noms propres, aucun dictionnaire ne serait assez exhaustif. Par ailleurs, faire appel à une technique endogène, qui par définition tire toutes les informations nécessaires du corpus à traiter, permet de passer outre à la phase préalable de constitution manuelle de données d'entraînement, extrêmement onéreuse. Enfin, une telle approche permet d'avoir un traitement parfaitement adapté au corpus traité, quel que soit le domaine : un système endogène est forcément adapté au corpus qu'il traite puisque, par essence, les données qu'il utilise pour traiter le corpus sont issues du corpus lui-même.

L'appariement (Figure 1, étape 3) s'effectue à partir des noms d'organisations pré-normalisés obtenus à la fin de la phase 2 (Figure 1, état c). En l'occurrence, pour notre exemple, la forme :

4 c) *Univ Kentucky Medical Center*

sera comparée à tous les autres noms pré-normalisés. Le système testera la fréquence de toutes les sous-séquences issues du segment 4 c) dans la base. Une sous-séquence est une suite de mots graphiques extraite du segment testé. Le découpage en sous-séquences s'exécute à partir de l'amorce de niveau supérieur, auquel on ajoute un mot pour la première sous-séquence. Le cas échéant, les premiers mots sélectionnés sont ceux qui sont situés à gauche de l'amorce, parce que la plupart des données problématiques à ce stade sont des données en anglais, et que la qualification des noms dans cette langue se fait souvent de droite à gauche, et non de gauche à droite comme pour le français. La taille de la fenêtre ainsi créée est incrémentée de 1 pour chaque nouvelle sous-séquence. Notons qu'en 4 c), l'amorce de niveau supérieur est *University of*, et non *University* : le mot outil *of* est considéré comme faisant partie de l'amorce. Le test des sous-séquences s'arrête lorsque le système détecte une amorce de niveau inférieur, puisque nous savons déjà que le nom d'organisation final ne peut contenir qu'une seule amorce. Pour 4 c), les sous-séquences testées sont donc les suivantes :

4 c.1) *University of Kentucky*

4 c.2) *University of Kentucky Medical*

Le système ne va pas plus loin, puisqu'il rencontre après *Medical* la seconde amorce, *Center*, qui n'appartient évidemment pas au nom d'organisation le plus cohérent.

Pour chacune de ces sous-séquences, le système calcule donc la fréquence dans l'intégralité des formes intermédiaires du corpus.

Nous partons de l'hypothèse selon laquelle la sous-séquence la plus fréquente est le nom d'organisation le plus cohérent.

A partir de notre corpus de plusieurs millions de noms d'organisations, nous avons relevé les résultats suivants :

11. *University of Kentucky* 71 occurrences
12. *University of Kentucky Medical* 9 occurrences

De fait, en sélectionnant la sous-séquence la plus fréquente, nous sommes en mesure de détecter le nom d'organisation le plus cohérent, soit :

11. *University of Kentucky*

Pour terminer, il ne reste qu'à procéder de nouveau à l'étape de réécriture, qui va en l'occurrence supprimer les mots outils, et abrégé *University* :

11. *University of Kentucky*
11') *Univ Kentucky*

Nous trouverons donc le nom de cette organisation dans les suggestions soumises à l'utilisateur en *d* (Figure 1). Ainsi, comme pour les résultats des calculs de distance d'édition, les normalisations réalisées par extraction endogène sont validées par l'utilisateur (étape 3' du schéma), et chaque décision est mise en mémoire. L'ensemble des noms normalisés est représenté en *e* sur notre schéma à la page 11.

Nous avons détaillé les différentes étapes de notre processus : tandis que la première phase est basée sur des règles intra-entité, l'étape de l'appariement par mesures de similarité et celle de la normalisation endogène tirent leurs informations de comparaisons inter-entités. Tous les résultats de ces comparaisons sont soumis à la validation de l'utilisateur expert, de manière à obtenir des données non bruitées et fiables et utilisables pour

toute application. Nous avons soumis les noms d'organisations de la première phase à une évaluation, de manière à établir le plus objectivement possible la qualité de cette pré-normalisation.

3. Évaluation

3.1. *Protocole d'évaluation*

Afin de mesurer la qualité de nos premiers résultats de manière fiable, nous avons établi un protocole d'évaluation. Rappelons que l'ensemble de notre système est un processus interactif, qui appelle une participation active de l'utilisateur. Ce caractère interactif a forcément un impact sur l'évaluation du système complet, puisque nous ne pouvons nous contenter dans ce cas de mesures telles que la précision et le rappel, fondées sur le décompte du bruit et du silence. D'autres paramètres sont à prendre en considération, par exemple le taux de compétence et le taux d'efficacité, qui consistent principalement à mesurer le nombre de corrections devant être apportées par l'utilisateur (Luzzati 1996). Nous pouvons également faire appel aux notions de l'ergonomie, qui décrivent la façon dont les utilisateurs perçoivent l'information qui leur est présentée (Bellies 2002).

La difficulté posée par l'interactivité du système est moindre pour le moment, puisque notre évaluation n'a porté que sur la phase de pré-normalisation à base de règles, pour laquelle nous ne faisons pas intervenir l'utilisateur. Cependant, dans la mesure où les attendus ont été définis en collaboration avec les utilisateurs experts, et où les attentes individuelles pouvaient varier, une certaine part de subjectivité est inhérente aux attendus mêmes. Malgré tout, nous avons objectivé ces attendus au maximum, en fixant pour chaque type d'entité nommée une forme standard à atteindre qui réponde au mieux aux besoins des utilisateurs. L'évaluation a donc consisté à comparer les noms obtenus avec ces attendus, à l'aide des mesures traditionnellement utilisées pour les systèmes de traitement automatique des langues, par exemple dans les conférences Text REtrieval Conference (TREC) ou Message Understanding Conference (MUC). Nous avons choisi de calculer le bruit et le silence obtenus pour un échantillon de 1000 noms d'organisations bruts, ainsi qu'un certain nombre de mesures que nous pouvons qualifier d'intermédiaires, et que nous décrivons en détail ci-dessous.

Corpus d'évaluation

Nous avons équilibré notre corpus d'évaluation de la manière suivante :

- La moitié des 1000 noms sont des noms de déposants de brevets, tandis que les autres sont des auteurs d'articles;
- Les publications proviennent de tous les domaines que nous avons signalés, avec une dominante médecine et biochimie, parce que l'une de nos sources principales est spécialisée dans ces domaines.

Élaboration des attendus

Nous avons mentionné que les attendus avaient été établis en collaboration avec des utilisateurs experts du domaine. Pour chaque type d'entité nommée, et en fonction des besoins des utilisateurs, nous avons défini un modèle standard auquel l'entité normalisée doit correspondre pour être jugée correcte. Nous présentons quelques exemples de normes :

- pour les entreprises : *Nom_de_l_entreprise* (sans identifiant de société nationale) :
- *Mitsubishi*
- pour les universités : *Univ + Nom_de_l_université* :
- *Univ Toulouse II*
- pour les instituts : *Inst + Nom_de_l_institut* ou *Nom_de_l_institut + Inst* :
- *Riverside Research Inst*
- pour les laboratoires : *Lab + Nom_du_laboratoire* ou *Nom_du_laboratoire + Lab* :
- *Lab Hygiene Wisconsin*

A côté de ces formes standard, le type de chaque organisation (entreprise, institution publique, ou particulier) et son pays d'origine sont également intégrés dans l'évaluation.

Détail des mesures

Nous présentons les divers critères qui nous permettent de réaliser une évaluation précise de nos résultats sur les noms, et de savoir où sont les

difficultés majeures. Nous ne nous sommes pas limités à la quantification des bruits et des silences, mais nous avons opté pour une typologie plus fine :

Tableau 1 : Critères d'évaluation de la normalisation

Intitulé de l'évaluation	Critères de sélection
Correct	Forme attendue, complète et sans parasite
Correct mais mots en désordre	Les mots extraits sont les bons, mais pas dans le bon ordre
Partiellement correct avec bruit	Forme attendue + éléments non pertinents
Partiellement correct avec silence	Forme attendue – un ou plusieurs éléments pertinents
Partiellement correct avec bruit et silence	Forme attendue + éléments non pertinents – un ou plusieurs éléments pertinents
Incorrect avec bruit	Le nom ramené n'est pas celui attendu et le système a ajouté du bruit. Ex. : <i>Institut de recherche contre le cancer</i> au lieu de <i>Université de Toulouse III</i>
Incorrect avec silence	Le système n'a effectué aucune normalisation, et le nom brut est incorrect

En ce qui concerne le type d'organisation et le pays d'origine, les critères d'évaluation sont plus restreints :

- le type ou le pays n'a pas été ramené, et c'est un problème de rappel ;
- le type ou le pays ramené n'est pas le bon, et c'est un problème de précision ;
- le type ou le pays a été correctement extrait.

Pour systématiser au maximum la tâche d'évaluation, nous avons créé une interface permettant de procéder à la vérification de chaque nom de l'échantillon et de traiter les données. De cette manière, les calculs sont automatisés et moins soumis à un risque d'erreur.

L'évaluateur vérifie chaque nom normalisé à partir de cette interface, en se fondant à la fois sur le nom initial et le nom normalisé en colonnes 2 et 3, et sur les modèles standard élaborés (*cf. supra*). Il doit cocher les

ID	ORIGINAL NAME	NAME NORMALISE	TYPE D'ORGANISATION	PAYS	SOURCE		
UNIV WASHINGTON							
9362	UNIV WASHINGTON	☒ Correct	☒ Correct ☐ Faux	☐ Correct ☐ Faux	2		
		☐ Partiellement correct avec bruit				☐ Manquant	☒ Manquant
		☐ Partiellement correct avec silence					
		☐ Partiellement correct bruit + silence ☐ Incorrect avec bruit ☐ Incorrect avec silence					
SIEMENS							
9968	SIEMENS AG	☒ Correct	☒ Correct ☐ Faux	☐ Correct ☐ Faux	2		
		☐ Partiellement correct avec bruit				☐ Manquant	☒ Manquant
		☐ Partiellement correct avec silence					
		☐ Partiellement correct bruit + silence ☐ Incorrect avec bruit ☐ Incorrect avec silence					
UNIV JOHNS HOPKINS							
10399	UNIV JOHNS HOPKINS	☒ Correct	☒ Correct ☐ Faux	☐ Correct ☐ Faux	2		
		☐ Partiellement correct avec bruit				☐ Manquant	☒ Manquant
		☐ Partiellement correct avec silence					
		☐ Partiellement correct bruit + silence ☐ Incorrect avec bruit ☐ Incorrect avec silence					
KONINKL PHILIPS ELECTRONICS							
11195	KONINKL PHILIPS ELECTRONICS NV	☒ Correct	☐ Correct ☐ Faux	☐ Correct ☐ Faux	2		
		☐ Partiellement correct avec bruit				☐ Manquant	☐ Manquant
		☐ Partiellement correct avec silence					
		☐ Partiellement correct bruit + silence ☐ Incorrect avec bruit ☐ Incorrect avec silence					
SIEMENS							
12520	SIEMENS AKTIENGESSELLSCHAFT	☒ Correct	☐ Correct ☐ Faux	☐ Correct ☐ Faux	2		
		☐ Partiellement correct avec bruit				☐ Manquant	☐ Manquant
		☐ Partiellement correct avec silence					
		☐ Partiellement correct bruit + silence ☐ Incorrect avec bruit ☐ Incorrect avec silence					
KARLSRUHE FORSCHZENT							
			non renseigné	UNKNOWN			

Figure 2 : Interface d'évaluation des noms normalisés

cases correspondant au cas de figure qui lui est présenté pour déterminer s'il est correct, et ce pour l'ensemble des noms d'organisations sur lesquels le système de normalisation a été appliqué.

3.2. Qualité de la normalisation obtenue

Nous présentons les taux de normalisation correcte pour les 1000 noms d'organisations de l'échantillon, en fonction du type de documents. Nous cumulon ensuite les résultats de la normalisation du nom avec l'extraction du type, puis avec l'extraction du type et du pays :

Signalons que sur ces 1000 noms d'organisations, 136 avaient déjà la forme attendue avant que le processus de normalisation ne soit appliqué. La normalisation est donc une étape indispensable, puisque 86,4 % des noms bruts ont une forme différente de celle que nous attendons.

Globalement, nous obtenons des résultats relativement satisfaisants pour la normalisation des noms, puisque sur les 1000 noms d'organisations, 840 ont la forme standard attendue (Tableau 2). Si nous retirons les 136 formes déjà correctes, qui ne sont pas modifiées par le système, 704 noms ont été correctement normalisés d'après nos normes.

Tableau 2 : Organisations correctement normalisées et typées

Organisations correctement normalisées	Nom	Type	Pays	Nom + Type	Nom + Type + Pays
Brevets	94,4 % (472/500)	73,4 % (367/500)	26,2 % (131/500)	70,4 % (352/500)	25,8 % (129/500)
Articles scientifiques	73,6 % (368/500)	94,2 % (471/500)	98,6 % (493/500)	71,6 % (358/500)	70,8 % (354/500)
Total	84 % (840/1000)	83,8 % (838/1000)	62,4 % (624/1000)	71 % (710/1000)	48,3 % (483/1000)

Tableau 3 : Répartition des erreurs du système sur les noms, par type de document

Répartition des erreurs	Correct, mais mots en désordre	Partielle-ment correct avec bruit	Partielle-ment correct avec silence	Incorrect avec bruit	Incorrect avec silence	Nombre total d'erreurs par type de document
Brevets	0 %	3,8 %	1,8 %	0	0	5,6 %
Articles	0,3 %	19,5 %	1,4 %	3,5 %	1,7 %	26,4 %

Dans tous les cas, lorsque les types sont jugés non pertinents, il s'agit d'un problème de silence, et jamais d'un manque de précision : soit l'information est présente et correctement identifiée, soit elle n'apparaît pas dans le nom d'organisation brut. Nous obtenons le même résultat pour l'information pays. La disparité de résultats à cet égard entre les brevets et les articles s'explique par le fait que les noms des auteurs d'articles contiennent de manière générale beaucoup plus d'informations que les noms des déposants de brevets.

Notons qu'il existe une différence importante entre les résultats des brevets et ceux des articles scientifiques en ce qui concerne la normalisation des noms hors extraction du type et du pays : alors que le nom des déposants de brevets est correctement normalisé dans 94,4 % des cas, seuls 73,6 % des noms d'auteurs d'articles correspondent aux normes établies. Le tableau de répartition des erreurs par type de document nous permet d'interpréter cet écart de manière plus précise : 19,8 % du nombre total de noms d'auteurs d'articles sont mal normalisés en raison d'un manque partiel de précision, parce qu'il contient à la fois toute

l'information pertinente, et un certain nombre d'éléments non pertinents. C'est le cas pour seulement 3,8 % des noms de déposants de brevets. La plupart des déposants de brevets sont des entreprises, tandis que la plupart des auteurs d'articles sont des institutions publiques. Comme nous l'avons signalé, les noms d'institutions publiques sont généralement beaucoup plus longs que ceux des entreprises, et ils contiennent beaucoup d'informations à éliminer. De fait, ils sont beaucoup plus difficiles à normaliser.

Le faible taux d'erreurs attribuables au silence montre que notre traitement permet, sinon une parfaite exhaustivité, mais du moins une couverture assez large. Nous pensons arriver à augmenter encore ce score de rappel en réajustant les lexiques, et particulièrement le lexique d'amorces.

La typologie d'évaluation des erreurs étant plus fine pour les noms que le bruit et le silence, il ne serait pas informatif de calculer rappel et précision. Par ailleurs, une étude est en cours afin de mesurer les taux de compétence et d'efficacité, et cette étude exige un suivi des utilisateurs.

3.3. Retombées pour la société TecKnowMetrix

Dans le cadre de l'activité de TecKnowMetrix, la possibilité d'automatiser au maximum le travail de normalisation comporte un double avantage : le gain de temps et l'homogénéité des normalisations. Cela permet d'abord aux utilisateurs, qui procédaient jusqu'alors de façon entièrement manuelle à la normalisation et pour qui il s'agit d'une tâche secondaire et laborieuse, de gagner un temps précieux qu'ils peuvent consacrer à leur principal métier. Ensuite, le risque d'erreur lié à la réalisation manuelle d'une telle tâche est réduit. On peut donc effectuer des traitements efficaces à partir de données normalisées fiables et non bruitées. La qualité de ces noms normalisés a donc une incidence sur la qualité des données des analyses ultérieures.

Notre système de normalisation permet ainsi de réduire le coût du temps d'exécution de cette tâche. Par ailleurs, en raison du caractère interactif du processus, il produit des données de qualité, puisque l'utilisateur peut intervenir lorsque surviennent des normalisations difficiles ou ambiguës pour le système. Enfin, le stockage des informations fournies par l'utilisateur lors de ses interventions et la réutilisation des données déjà normalisées pour toute nouvelle normalisation donnent un aspect itératif à l'ensemble du système. On peut ainsi réduire encore le temps d'intervention de l'expert à chaque nouvelle application du traitement.

3.4. *Retombées pour la recherche d'information*

Dans le domaine de la recherche d'information, le principal avantage d'une telle normalisation est le gain de rappel : si l'ensemble des variantes d'une entité nommée sont identifiées et ramenées à une même forme graphique, le système de recherche d'information pourra retourner un plus grand nombre de réponses pertinentes. Dans le cas contraire, une partie serait ignorée si aucun traitement préalable n'était appliqué. Cet avantage a déjà été démontré par plusieurs études récentes dans le domaine, comme Jijkoun et al. (2008) et Khalid et al. (2008).

Dans ce contexte, il est possible, grâce au caractère interactif de notre processus, de personnaliser le système en fonction des besoins et des attentes de chaque utilisateur. Le fait que le système soit endogène et qu'il utilise peu de ressources externes en fait un outil exploitable dans de nombreux domaines.

Les avantages d'une normalisation des entités nommées pour la recherche d'information se retrouvent également pour l'extraction d'information et les systèmes de question-réponse : plus le nombre variations linguistiques gérées est important, plus les systèmes sont performants.

3.5. *Apports pour d'autres applications*

Pour la constitution de ressources linguistiques en général, et pour la construction d'ontologies ou de ressources termino-ontologiques en particulier, la normalisation est indispensable. Un système capable d'identifier et de normaliser tous les noms référant à un même objet du monde permet tout d'abord d'avoir des données standardisées. En outre, cela évite d'intégrer des doublons en tant que noeuds du réseau et qui rendraient l'intégralité de la structure du réseau incorrecte, ce qui engendrerait des problèmes de cohérence.

Enfin, l'élimination de doublons au sein de bases de données par la normalisation a forcément des retombées importantes, dans la mesure où cela garantit la qualité et la non-redondance des informations.

Conclusion

A travers cet article, nous avons abordé le problème complexe de la normalisation des entités nommées, problème situé en amont de diffé-

rentes tâches de traitement de l'information. Les retombées impliquées par une normalisation précise des entités nommées sont très importantes lorsque les tâches de recherche d'information et d'analyse bibliométrique doivent être le plus précises possible, comme c'est le cas pour la société TecKnowMetrix et aussi pour tous les acteurs du domaine du traitement de l'information. En proposant un processus de normalisation fondé à la fois sur des traitements à plusieurs niveaux textuels (intra-entité et inter-entités) et prenant en considération les validations de l'utilisateur, nous avons apporté une contribution conjointe à des domaines comme ceux de la recherche d'information et de la représentation de connaissances. La nécessité de la tâche de normalisation est liée à la prépondérance des entités nommées dans les textes. Sans une normalisation pertinente de ces entités, il est difficile de les identifier dans les textes et de les regrouper.

Les entités nommées constituent des points d'accès aux textes. Ces derniers les mettent très souvent en relation entre elles ou avec d'autres segments textuels qui, entre autres, les caractérisent, les décrivent, et les évaluent. Représenter ces relations dans une ressource termino-ontologique permettrait d'isoler un grand nombre de renseignements sur les entités nommées et les textes où elles figurent. Nos travaux de recherche et développement sont donc orientés en ce sens. Nous cherchons à construire une telle ressource pour chacun de nos domaines de travail. La visualisation et la navigation dans cette ressource constitueraient un accès guidé au corpus et nous permettraient sûrement de peaufiner notre analyse.

À titre d'exemple, les trois auteurs du présent article créent un lien termino-ontologique entre une université et une entreprise privée (LIDILEM et TecKnowMetrix). Par leurs autres publications, ils créent d'autres liens, alimentant ainsi un réseau complexe entre entités nommées elles-mêmes associées à des documents qui décrivent ces liens.

La construction de cette ressource termino-ontologique ne serait pas possible sans une normalisation des entités nommées. Cette tâche de normalisation doit ainsi être située en amont de tout traitement d'information textuelle, afin de garantir sa pertinence et de fournir ainsi aux utilisateurs le meilleur accès possible au contenu des textes.

Note

1. Mitsubishi est une marque enregistrée de Mitsubishi Corporation.

Références

- Alphonse, E., S. Aubin, P. Bessires, G. Bisson, T. Hamon, S. Lagarrigue, A. Nazarenko, C. Ndellec, M. Ould Abdel Vetah, T. Poibeau, et D. Weissenbacher. 2004. Extraction d'information applique au domaine biomdical—apprentissage et traitement automatique de la langue, Actes de la *Confrence Internationale de Fouille de texte*.
- Aussenac-Gilles, N. 2007. *Projet DaFOE4App—Dossier A.0 / Document A.0.1—tat de l'art et tude des besoins pour une plateforme de construction d'ontologies*, rapport de contrat, IRIT/RT—2007-2—FR, IRIT.
- Bellies, L. 2002. La conception : processus d'laboration et d'valuation de reprsentations pour l'action, thse de doctorat en ergonomie, Paris, cole Pratique des Hautes tudes.
- Bourigault, D. 1993. An Endogenous Corpus Based Method for Structural Noun Phrase Disambiguation, actes de la *Conference of the European Chapter of ACL (EACL)*, 81–86.
- Bourigault, D., N. Aussenac-Gilles, et J. Charlet. 2004. Construction de ressources terminologiques ou ontologiques partir de textes : Un cadre unificateur pour trois tudes de cas. *Revue d'intelligence artificielle* 18, 87–110.
- Bourigault, D., et C. Frrot. 2006. Acquisition et valuation sur corpus de proprits de sous-catgorisation syntaxique, *TAL* 47, 141–54.
- Elmagarmid, A.K., P.G. Ipeirotis, et V.S. Verykios. 2007. Duplicate Record Detection : A Survey, *IEEE Transactions on Knowledge and Data Engineering* 19, 1–16.
- Jijkoun, V., M.A. Khalid, M. Marx, et M. de Rijke. 2008. Named Entity Normalization in User Generated Content, actes de *SIGIR 2008—Workshop on Analytics for Noisy Unstructured Text Data*.
- Khalid, M.A., V. Jijkoun, et M. de Rijke. 2008. The Impact of Named Entity Normalization on Information Retrieval for Question Answering. *LNCS* 4956.
- Labb, C., et D. Labb. 2006. La diachronie dans le discours politique : Le gnral de Gaulle, actes des *Nouvelles journes de l'ERLA*.
- Lebarb, T. 2007. LexTract : Extraction semi-automatique de termes porte juridique, actes des *Journes de la linguistique de Corpus*, 197–205.
- Levenshtein, V. 1966. Binary codes capable of correcting deletions, insertions and reversals, *Soviet Physics Doklady* 10, 707–10.
- Luzzati, D. 1996. Le dialogue verbal homme-machine, tudes de cas, Paris, Masson.
- Mann, G. 2002. Fine-Grained Proper Noun Ontologies for Question Answering, actes de *SemaNet'02 : Building and Using Semantic Networks*.
- Poibeau, T. 2001. Deconstructing Harry, une valuation des systmes de reprage d'entits nommes, *Revue de la socit d'lectronique, d'lectricit et de traitement de l'information*.
- Poibeau, T. 2003. *Extraction automatique d'information : du texte brut au web smantique*, Paris, Herms.
- Tesnire, L. 1969. *lments de syntaxe structurale*, Paris, Klincksieck.
- Vergne, J. 2004. Dcouverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource, actes des *JADT 2004* 2, 1158–64.